REVIEW

Structural genomics: Computational methods for structure analysis

SHARON GOLDSMITH-FISCHMAN AND BARRY HONIG

Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA

Abstract

The success of structural genomics initiatives requires the development and application of tools for structure analysis, prediction, and annotation. In this paper we review recent developments in these areas; specifically structure alignment, the detection of remote homologs and analogs, homology modeling and the use of structures to predict function. We also discuss various rationales for structural genomics initiatives. These include the structure-based clustering of sequence space and genome-wide function assignment. It is also argued that structural genomics can be integrated into more traditional biological research if specific biological questions are included in target selection strategies.

Keywords: Structural genomics; homology modeling; structure alignments; functional analysis.

Structural genomics is a term that refers to high-throughput three-dimensional structure determination and analysis of biological macromolecules, at this stage primarily individual protein domains. The determination of the three-dimensional structures of proteins has for many years come under the classification of "curiosity" or "hypothesis driven" research. Structures were generally determined because they could be expected to teach us something new about a biological problem; for example, the details of an enzyme mechanism, the nature of a molecular recognition process, or the energetic basis of energy transduction processes. An important spinoff of structural biology has been the discovery of new relationships between amino acid sequences and protein structures, and among different protein structures. New computational tools have been developed to exploit the information that has become available, and many remarkable and unexpected relationships have been uncovered. Concepts such as protein family, fold, and superfamily have been introduced (Orengo et al. 1997; Hubbard et al.

1999), and detailed taxonomies have been developed that help us understand the complex three-dimensional shapes of proteins. Structural genomics represents a new direction in structural biology in that it is based on the goal of determining as many structures as possible, even in advance of a well-defined biological question. Nevertheless, the field is ultimately "curiosity driven" but the questions being asked now relate to the discovery of complex relationships in sequence and structure space and, ultimately, to a deeper understanding of many biological problems once these relationships are understood.

It should be recognized that in the short run, structural genomics will not lead to the determination of the large macromolecular structures and complexes that have been the hallmark of the breathtaking advances in structural biology in recent years. Indeed, the experimental focus has generally been on technological advances in expression, purification, crystallization, and structure determination, whereas the structures themselves need not be particularly interesting in terms of the new biology they reveal. The stated goals of structural genomics have reflected this reality. One goal that was expressed at an early stage was the determination of a representative for each protein fold, thus providing complete coverage of "fold space." However, there is a great deal of ambiguity in the definition of a fold,

and in fact, an argument can be made that fold space should be viewed as continuous (Shindyalov and Bourne 2000; Yang and Honig 2000a; Harrison et al. 2002). A more precise expression of the goals of structural genomics is the experimental determination of enough structures so that all other structures can be built with homology modeling (Vitkup et al. 2001; Chance et al. 2002). The figure of 30% sequence identity is often used as the cutoff, above which a homology model is viewed as meaningful. Based on this criterion, it has been estimated that about 16,000 structures need to be determined (Vitkup et al. 2001). It is perhaps remarkable, and for a computational biologist quite satisfying, that an experimental strategy is based on the ability to construct a model. Functional considerations are also being used as criteria in target selection, as discussed in a recent article in this journal from the New York Structural Genomics Research Consortium (Chance et al. 2002). An additional strategy, based on the functional information derivable from homology models, will be introduced below.

Although the goals of structural genomics are varied, there is no question that the coming years will see a major increase in the number of proteins whose three-dimensional structures are known. The status of structural genomics targets, including information relating to gene expression, protein purification, and structure determination, is curated and made publicly available at targetdb.pdb.org. How should the newly determined structures be used? As part of NESG (Northeast Structural Genomics Consortium; www.nesg.org; Bertone et al. 2001) we are attempting to derive maximum information from each new structure that is determined by our consortium. Our goals are to assign function or to suggest functional hypotheses for each new structure and to update sequence–structure relationships on a continuous basis. The availability of new structures allows us to test existing methods and to consider what new computational tools might be developed so as to better exploit the continuous flow of new structural information. The goal of this article is to discuss the approaches currently being used to analyze new structures and, in a number of cases, to point to new directions. Specifically we summarize (1) structural relationships among proteins, (2) methods that combine sequence and structural information to derive new relationships between distantly related proteins, (3) protein structure prediction by homology, and (4) structure-based assignment of protein function. As will be discussed, there is a clear synergistic relationship between computational methods and target selection in structural genomics. Computational methods generally become increasingly effective when the data set of protein structures upon which they are based increases. In parallel, the criteria used to choose the structures to be determined experimentally will benefit from improvements in the sensitivity and precision of methods used to define sequence, structure, and functional relationships between proteins.

Structure alignments

Folded proteins are constructed from a small number of secondary structure elements (SSEs) whose mutual organization in space creates distinct and somewhat striking three-dimensional patterns. These were visually characterized and classified in a seminal paper by Jane Richardson, who also emphasized their esthetic qualities (Richardson 1981). The visual classification of protein structures and the relationships that can be detected in this way are embodied in the SCOP database (Hubbard et al. 1999). SCOP classifies most globular proteins into class (Levitt and Chothia 1976), fold, superfamily, and family; these terms have been widely adopted. Fold refers primarily to the organization of SSEs in space, while superfamily refers to proteins with the same fold and a related function. In addition to SCOP, another widely used classification scheme is the CATH database of Thornton, Orengo et al. (1997). CATH is partially automatic and makes use of the SSAP program (Orengo and Taylor 1996) to derive structural relationships. As is the case for SCOP, CATH also involves manual intervention, particularly to discriminate between groups of proteins that have a similar fold (analogs) and those that have a common fold and a functional relationship (homologs, or superfamily members in the SCOP terminology).

SCOP and CATH are extremely valuable resources that have been used to derive structure/function relationships between proteins. SCOP by its very nature does not provide objective measures by which to describe structural relationships, whereas CATH, although including such a score (SSAP), is not based exclusively on structural relationships. A variety of objective measures are available from the many structure superposition algorithms that have been introduced in recent years (Swindells et al. 1998), all of which are based on one or more geometric criteria that are use to characterize structural similarity. For example, DALI (Holm and Sander 1993) measures structural similarity based on Cα contact distances while others algorithms such as SSAP (Orengo and Taylor 1996), VAST (Madej et al. 1995), and PrISM (Yang and Honig 2000a) make use of secondary structure information. The CE algorithm (Shindyalov and Bourne 1998) utilizes aligned fragment pairs of a given length; the program MAMMOTH (Ortiz et al. 2002) performs sequence- and secondary structure-independent structural alignments by detecting subsets of aligned pairs; the MUSTA program (Leibowitz et al. 2001) performs multiple structure alignments implementing an alignment algorithm based on a geometric hashing technique that is sequence-independent (Nussinov and Wolfson 1991). Structural similarity is often measured in terms of RMSD, but this can be ambiguous, as the definition of topologically equivalent residues to be used in the calculation of an RMSD is not always clear. The problem becomes more extreme for distantly related proteins. In addition, RMSD

measures are clearly more meaningful as the length of the aligned segments increases, a problem that has recently been addressed by Carugo and Pongor (2001), who introduced a normalized RMSD score. Structural similarity is often reported in terms of a Z-score, which measures a deviation from a database average. Z-scores determined with different algorithms have no direct relationship but, in the context of a widely used program such as DALI, provide an extremely meaningful measure of structural relationships. PrISM, a program developed in our lab (Yang and Honig 2000a), defines a protein structural distance (PSD) that is a composite score including a contribution from both an RMSD term and a term that accounts for the spatial arrangement of SSEs. In this way it attempts to measure similarities even for distantly related proteins.

This brief summary illustrates the underlying reality that there is no unique and objective measure for defining structural relationships between proteins. Nor, as pointed out by Godzik, is there unique way of aligning two proteins (Godzik 1996). On the other hand, most algorithms yield similar conclusions as to what is similar and what is not. Significant discrepancies do, however, arise, and it may be worthwhile utilizing multiple approaches in a particular application.

The identification of relationships between proteins based on structural alignments has a number of goals. The identification of homologs permits functional inferences so clearly revealed in the SCOP and CATH databases. In contrast, identifying structural analogs provides no obvious functional information but may be extremely useful if one is interested in common sequence features that cause a protein to adopt a particular topology (Yang and Honig 2000b). It is interesting in this regard that it is difficult to distinguish homologs from analogs on purely structural terms (Rost 1997; Russell et al. 1998; Yang and Honig 2000b). This difficulty reflects the fact that sequence features that determine structure are often distinct from those that determine function. Indeed, it is often the case that there is no detectable sequence conservation pattern common to structures with the same fold. Thus, when a well-defined sequence pattern is observed it is generally an indication of an evolutionary relationship, suggestive of a functional relationship, rather than resulting from the sequence requirements for a particular fold (Yang and Honig 2000c).

Sequence- and structure-based remote homolog and analog detection

Because the detection of functional homologs is the central approach currently used to assign function to specific genes, increasing the sensitivity of homolog detection is a problem of central importance. Pairwise sequence search methods such as BLAST and FASTA (Pearson and Lipman 1988; Altschul et al. 1997) can reliably identify homologs down to

about the 30% sequence identity level. More recently, multiple sequence profile methods such as PSI-Blast (Altschul et al. 1997) and Hidden Markov Models (HMM; Krogh et al. 1994; Eddy 1995, 1996; Karplus et al. 1998) have significantly enhanced our ability to detect remote homologs beyond what is possible with pairwise sequence methods. Databases such as Pfam (Bateman et al. 2002) and SMART (Schultz et al. 2000) contain multiple sequence alignments of individual protein and domain families. Other sequence derived database include PROSITE (Falquet et al. 2002), which contains profiles generated from multiple sequence alignments; EMOTIF (Huang and Brutlag 2001), which contains the sequence alignments from the BLOCKS+ and PRINTS databases (Attwood and Beck 1994; Henikoff et al. 1999), and InterPro (Apweiler et al. 2000), a database combining information present in the PRINTS, PROSITE, Pfam, TIGERFAMs (Haft et al. 2001), and ProDom (Servant et al. 2002) databases and cross-referenced with BLOCKS. These databases can be searched for functional relationships using Web-based servers.

Pure sequence-based methods have inherent statistical limits. The use of structural information has been shown to increase both the detection sensitivity and alignment accuracy once a relationship has been detected. Structure-based 3D–1D sequence profiles (Bowie et al. 1991) and fold recognition methods (Jones et al. 1992) have been in use for some time (Jones 1997) and will not be reviewed here. Recent advances in both of these approaches have involved methods that combine sequence, structural, and functional information (Johnson et al. 1993; Elofsson et al. 1996; Fischer et al. 1996; Rice et al. 1997; Rost 1997; Jaroszewski et al. 1998; Hargbo and Elofsson 1999; Jones et al. 1999; Kelley et al. 2000; Panchenko et al. 2000; Kolinski et al. 2001; Shi et al. 2001; Reva et al. 2002).

The use of multiple structure alignments to derive multiple sequence alignments offers an alternative approach to the mapping of structural information onto sequence (Mizuguchi et al. 1998; Matsuo and Bryant 1999; Yang and Honig 2000c; Reddy et al. 2001). Sequence profiles generated from multiple structure alignments have been used to identify homologous core structures (Matsuo and Bryant 1999), to identify evolutionarily conserved residues (Mirny and Shakhnovich 1999; Yang and Honig 2000c; Reddy et al. 2001), and to derive structure-based substitution matrices (Ogata et al. 1998; Prlic et al. 2000; Blake and Cohen 2001). On the other hand, although the use of multiple structure information can improve alignment quality (Panchenko et al. 2000; Shi et al. 2001), sequence profiles generated entirely from multiple structure alignments alone do not perform particularly well (Panchenko et al. 2000; Gough and Chothia 2002; Griffiths-Jones and Bateman 2002). These papers have shown that profiles constructed with the individual family members as seeds performed better than those constructed with a single combined alignment of an entire

family. Recently, multiple structure alignments have been combined with multiple sequence alignments to enhance both sequence search sensitivity (Kelley et al. 2000) and alignment quality (Al-Lazikani et al. 2001). In these approaches, sequences for which structures are available are aligned with a multiple structure alignment. Each of these sequences is then used as a seed to align to related sequences using a multiple sequence alignment. The individual multiple sequence alignments are then merged through the multiple structure alignment to yield what we will term a sequence enhanced multiple structure alignment. However, even when sequence and structural information are combined, strong sequence signals can at times be lost when three-dimensional position specific scoring matrices (3D-PSSM) are used (Kelley et al. 2000).

Homology modeling

Homology or comparative modeling involves the prediction of the structure of a query sequence from the structures of one or more structural templates. The procedure involves the identification of possible templates that have a clear sequence relationship to the query, the assembly of the model, the prediction of regions of the structure that are likely to have different conformations than the templates (e.g., loops), and ultimately, the refinement of the structure in an attempt to account for inherent differences between the template and query structures. As mentioned above, homology modeling figures heavily as a rationale for structural genomics initiatives under the stated assumption that accurate models can be built for query sequences that have a greater than 30% sequence identity with their best template. Of course, the accuracy requirements for a homology model depend in large part on why the model is being built. For example, if one is using a model in structure-based drug design there is a clear need for a highly accurate description of the ligand binding site. In contrast, if an electrostatic pattern on a protein surface (Honig and Nicholls 1995) is of interest to help identify a binding interface with another protein, nucleic acid, or membrane, it may be possible to suffice with a less reliable model based on lowers levels of sequence identity to the template. Thus, the 30% rule is best used as a useful guideline rather than as a meaningful cutoff as to when a model should be viewed as reliable.

The quality of the alignment of the query to the template sequence is a major factor in determining the quality of homology models. This is one of the sources of the 30% rule, because alignment quality usually decreases dramatically below about 30% sequence identity. (A structural explanation for this observation has been offered by Chung and Subbiah, 1996). On the other hand, continuing improvements in profile-based sequence alignment methods have extended the range of sequence identities where it seems

appropriate to attempt the construction of a homology model. It is interesting in this regard that some of the homology modeling targets in the CASP4 (Venclovas 2001) and CASP5 experiments have no homologs in the PDB that have significant levels of sequence identity with the query. Rather, the assignment to the homology category is based on the availability of statistically significant Psi-Blast hits (Tramontano 1998). Advances in the accuracy of sequence alignments using structure-based profile methods such as those described above should result in continuing improvements in the quality of homology models and in an increase in the number of sequences for which a meaningful homology model can be built.

Once one or more templates have been identified and alignments have been decided, a decision must be made as to how to construct the model. A number of strategies have been adopted (Sanchez and Sali 1997). When one template is clearly preferable, the coordinates of the aligned residues in the query are simply superimposed on those of the template. When more than one template is available, it is possible to construct a model based on multiple templates where the coordinates of the query are required to satisfy spatial constraints defined by interatomic distances in each of the templates. This is the main strategy adopted in the widely used MODELLER program (Sali and Blundell 1993). A third option is to construct a composite model that is assembled from structural fragments taken from different templates. A new program, Nest, developed in our lab, allows the user to choose any one of these options or to allow the program to decide on its own which option to choose (http://trantor.bioc.columbia.edu/~xiang/jackal/index.html; Z. Xiang and B. Honig, in prep.).

Once a model has been constructed for the backbone, the conformations of side chains that are different between the query and template need to be predicted. The standard procedures is to sample rotamer libraries for each residue (Ponder and Richards 1987; Xiang and Honig 2001; Dunbrack 2002) and search for the combinations of side chain conformations with the lowest energy (Bower et al. 1997; Levitt et al. 1997). There are a large number of possible rotamer combinations for an entire protein, and this has led to the application of advanced sampling techniques to the problem (Lee and Subbiah 1991; Vasquez 1996; De Maeyer et al. 1997). These seem particularly important for problems of protein design where the actual sequence of the protein needs to be determined (Dahiyat and Mayo 1997; Gordon and Mayo 1999; Looger and Hellinga 2001). However, the combinatorial problem is much less severe than what was generally assumed if one wishes to predict the conformations of buried side chains in a given protein, presumably because the problem of packing the protein interior is so constrained by steric factors (Xiang and Honig 2001). If a large number of rotamers per residue are sampled, it is generally possible to find low energy conformations that

correspond quite closely to the native structure (Xiang and Honig 2001; Jacobson et al. 2002).

A second problem that has been addressed for many years is the prediction of the conformation of loops. These often correspond to regions of the query sequence that are different than those of the template so that the problem of loop prediction is an important element in homology modeling. Two general approaches have been applied to the prediction of loop conformation: database search and ab initio techniques. In the database search method, a library of segments derived from known protein structures is searched for conformations that fit the topological constraint of the loop stems. Loop candidates found in this way can then be evaluated by different criteria such as sequence relationships between the template and query segment or some measure of conformational energy. In some applications, such methods can be very powerful; for example, when canonical structures exist, as is the case for the hypervariable loops of antibodies (Chothia and Lesk 1987; Martin and Thornton 1996). However, in general, there is no guarantee that the correct loop conformation can be found in the PDB. Ab initio methods involve the generation of a large number of loop conformations, usually randomly, and their evaluation based on some sort of energy function (Bruccoleri and Karplus 1990; Rapp and Friesner 1999; Fiser et al. 2000). Recent advances in ab initio loop prediction (Fiser et al. 2000; Xiang et al. 2002) suggest that the approach will, in general, yield significantly more accurate predictions than fragment based methods. This is because it is possible to generate and evaluate far more conformations for a given loop than are available from fragments in the PDB.

Despite progress in sequence alignment, model building, and loop and side chain prediction, formidable problems still remain in homology modeling. Even the most accurate loop and side chain procedures only work well when the conformation of the backbone is accurately known, and this often is not be the case. Indeed, there are frequently differences in conformation between query and template, and unless these can be predicted, they pose an inherent limitation on all model-building procedures. This is another reason that prediction accuracy declines at low levels of sequence identity; as proteins diverge more in sequence, they tend to diverge more in structure (Yang and Honig 2000b). Thus, even if an alignment is perfect, the accuracy of the homology model will depend on the degree of structural similarity between template and query. Structural genomics initiatives will help solve this problem by providing more templates with significant levels of sequence identity to every protein whose structure is not known. In addition, continuing advances in the ability to evaluate conformational free energies offer the possibility of significant improvements in structure prediction.

Programs such as Verify 3D (Luthy et al. 1992) and Prosa II (Sippl 1993) that provide measures of protein stability are already widely used in evaluating the reliability of different models and in identifying regions of a structure that do not appear to be native-like. However, if a truly accurate method of evaluating relative conformational free energies were available the procedure of constructing a large number of models for each protein and choosing based on their relative stabilities would become far more effective. Conformational free energy "scoring functions" fall into two categories: statistical effective energy functions that are based on the observed properties of amino acids in known structures and physical effective energy functions that are based on a direct evaluation of the conformational free energy of a protein (Sippl 1995; Samudrala and Moult 1998; Simons et al. 1999; Lazaridis and Karplus 2000; Petrey and Honig 2000). Both approaches are generally capable of discriminating native conformations from incorrectly folded "decoys," but this does not guarantee that they are able to determine which of two partially incorrect structures is closest to the native conformation. Even if this was possible, the problem of beginning with a structure that differs from the native, and relaxing it in such a way so as to arrive at a near-native conformation, or at least a structure that is closer to the native than the initial conformation, still remains. Of course, the two problems are closely linked; the more accurate the evaluation of the conformational free energy, the more likely it is that some procedure that produces conformational change, such as molecular dynamics, will be able to improve on a model constructed from one or more templates. The goal of significantly improving a structure constructed from one or more templates poses a major computational and theoretical challenge that must be overcome if there is to be significant progress in homology modeling beyond what will result from the availability of an increasing number of structures.

Structure-based functional analysis

The functional annotation of newly determined structures (Teichmann et al. 2001) is important, not only for gaining biological insights, but for uncovering novel relationships between sequence, structure, and function. There are a number of ways in which structure determination can aid in the assignment of function. Some of these involve enhancements in widely used methods while others will require the development of totally new technologies that exploit three dimensional structures in novel ways. The assignment of function based on sequence homology is, of course, the most direct way of assigning function so that the advances in remote homolog detection based on structural information, as summarized above, will inevitably improve function assignment. Similarly, structural alignment methods allow the direct determination of relationships between proteins that are not evident from sequence analysis so that each new

structure has the potential of revealing new functional relationships (Zarembinski et al. 1998; Hwang et al. 1999; Volz 1999; Du et al. 2000).

A number of groups have investigated the correlation between sequence, structure, and functional similarity using E.C. numbers as a measure of functional relationships, including identifying the relationship between CATH or SCOP classification and E.C. number (Martin et al. 1998; Hegyi and Gerstein 1999; Devos and Valencia 2000; Pawlowski et al. 2000; Wilson et al. 2000; Todd et al. 2001; Rost 2002). Thornton and coworkers (Martin et al. 1998) report that enzyme function is not closely related to protein fold, indeed, it is well known that enzymes with very similar functions can have very different folds (the classic example being serine proteases such as subtilisin and trypsin). In contrast, a correlation was found between protein architecture and ligand type. Russell et al. have found that there are locations on groups of analogous proteins that show a tendency to bind substrates despite the absence of any evidence that these proteins are evolutionarily related (Russell et al. 1998). It is not clear that the existence of such "supersites" suggests the existence of a common ancestor or whether they simply reflect an underlying structural or physical-chemical property of a particular fold. Studies of this type raise intriguing questions whose answers may become more accessible as the number of available three-dimensional structures grows. These goals will be aided by efforts such as the Gene Ontology (GO) project, which provide a more detailed functional annotations of gene products than has been available in the past (Ashburner et al. 2000). GO classifies sequences into multiple categories, based on their specific function, biological process, and subcellular location, and should thus enable the detection of relationships that would not otherwise be evident.

Although the expectation is that many proteins solved as a result of structural genomics initiatives will not have identifiable sequence or structural homologs, it is clear that in many cases alignments of global structures will reveal relationships that aid in the assignment of function. However, there are many ways to define function, and as the desired definition becomes increasingly precise it becomes necessary to examine local motifs rather than global features. In parallel with the sequence motif databases that have been assembled (e.g., BLOCKS, PRINTS, PROSITE) a number of groups have developed methods to define three-dimensional motifs so that each new structure can be searched for the presence of a particular local pattern. For example, Wallace et al. (1996, 1997) have assembled a database of three-dimensional templates of active site residues. Their PROCAT database can be searched (http://www.biochem. ucl.ac.uk/bsm/PROCAT/PROCAT.html) to identify groups of residues in a query structure with orientations consistent with those in known active sites (Wallace et al. 1996). Kasuya and Thornton have shown that many PROSITE pat-

terns have a common three-dimensional structure that provides a basis for creating a library of functional templates (Kasuya and Thornton 1999). Skolnick et al. developed geometric and conformational descriptors of active site residues (Fetrow and Skolnick 1998; Fetrow et al. 1999). These "fuzzy functional forms" contain information from sequence conservation, and biochemical data, as well as describe geometric relationships between active site residues.

Another approach that has been used to identify functionally important regions involves the mapping of conserved sequence features on the protein surface. Methods in this category include Evolutionary Trace (Lichtarge et al. 1996), ConSurf (Armon et al. 2001), 3D-Cluster (Landgraf et al. 2001), and AL2CO (Pei and Grishin 2001). These approaches have been used to successfully identify clusters of specific ligand binding residues (Aloy et al. 2001; Armon et al. 2001; Pei and Grishin 2001; Lichtarge and Sowa 2002; Madabushi et al. 2002).

The physical and chemical nature of the protein surface offers an approach to the analysis of function that has only been partially exploited. Electrostatic, hydrophobicity, and geometric patterns such as those revealed in the GRASP program (Nicholls et al. 1991) have been used for some time to identify functionally important regions on protein surfaces. For example, patches of electrostatic potential are often indicators of a binding interface, usually to a molecular with a potential of opposite sign (Honig and Nicholls 1995). This is, however, not always the case; indeed, some interfaces appear to exploit electrostatic interactions to drive binding, while in others hydrophobic residues appear to be the dominant surface feature (Sheinerman and Honig 2002).

The quantitative description of protein surfaces is a problem that offers considerable potential in the structure-based analysis of function. In analogy with structural alignment methods that have been developed to identify common geometric features in the polypeptide backbone, it would be extremely valuable to be able to identify common surface features that are characteristic of a particular binding function. These might include descriptors of shape, electrostatic potential, hydrophobicity, and sequence conservation, mapped onto a protein surface rather onto individual amino acids. A number of efforts with this goal in mind have already been reported. These include patch analysis of (Jones and Thornton 1997), which analyzes surface features of patches of residues, and the GRASS (Nayal et al. 1999) and SPIN servers (www.trantor.bioc.columbia.edu), which an be used to identify functionally relevant residues in protein structures. Recently, Klebe and coworkers (Schmitt et al. 2002) reported a new method to recognize similar binding pockets on protein surfaces independent of any sequence or structural relationship these proteins might have. They describe a database (Cavbase) that contains active-site cavity descriptors that can be searched to help assign function to proteins of that may exhibit similar binding properties.

Discussion

In this review we have summarized (although not exhaustively) computational methods that are an integral component of current structural genomics initiatives. These methods can aid in the detection of novel relationships between proteins and in assigning function to individual proteins. Computational tools in this area can be expected to improve as additional data become available and as our understanding of the principles of protein structure and function continues to develop. Indeed, the advent of structural genomic initiatives is certain to spur the development of a host of new computational methods aimed at detecting new relationships between sequence, structure, and function.

Large-scale analysis is a common underlying theme in much current research effort in areas characterized by the "omics" suffix. Genome-wide structure prediction and function assignment are agreed upon goals as is the structure and function-based clustering of sequence space. These goals are clearly of great value, but they are only part of the picture. Much of biology still involves a focus on individual problems, and this is likely to remain the case for years to come. Thus, it becomes appropriate to ask how the vast quantities of new data can be used to address specific problems in detail rather than to provide a broad overview of genome-wide behavior. One approach is to develop a detailed structural description of entire protein families that is not accessible if the structures of just a few family members are available. It is important, for example, not only to know what family members have in common but also to understand how they are different. Biological specificity lies in the differences, for example, in the nucleotide sequence-specific DNA recognition of closely related transcription factors or in the differential phosphotyrosine containing peptide binding of different SH2 domains. Knowing the structures of all family members would clearly be useful in addressing questions such as these, but the spirit of many structural genomics initiatives is to avoid solving the structures of many closely related proteins.

The methods described in this review offer at least a partial solution to the problem. Continued progress in the development of sequence and structure alignment methods will increase the sensitivity of remote homolog detection and alignment accuracy will continue to improve. In parallel, homology modeling will become an increasingly more accurate procedure, and it will become possible to construct meaningful homology models for entire protein families. Such models will provide the basis for a more detailed analysis of function than has been available in the past and, in the hands of researchers interested in the biology of a particular problem, will provide powerful tools for the analysis of experimental data and for the design of new experiments (see, e.g., Murray and Honig 2002). The number of structures required to model all members of a par-

ticular family will not necessarily conform to the 30% rule mentioned above, but rather may be dictated by the properties of the family and by the particular question being asked. This protein family-based approach will still conform to the structural genomics approach of solving enough structures so that all others can be obtained from homology modeling, but the number of structures needed to be solved will be determined on a case by case basis. A target selection strategy with these ideas in mind may provide a link between structural genomics initiatives and more traditional research in structural biology.

References

Al-Lazikani, B., Sheinerman, F.B., and Honig, B. 2001. Combining multiple structure and sequence alignments to improve sequence detection and alignment: Application to the SH2 domains of Janus kinases. *Proc. Natl. Acad. Sci.* **98:** 14796–14801.

Aloy, P., Querol, E., Aviles, F.X., and Sternberg, M.J. 2001. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311:** 395–408.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2000. InterPro—An integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16:** 1145–1150.

Armon, A., Graur, D., and Ben-Tal, N. 2001. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307:** 447–463.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Attwood, T.K. and Beck, M.E. 1994. PRINTS—A protein motif fingerprint database. *Protein Eng.* **7:** 841–848.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., and Gerstein, M. 2001. SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* **29:** 2884–2298.

Blake, J.D. and Cohen, F.E. 2001. Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.* **307:** 721–735.

Bower, M.J., Cohen, F.E., and Dunbrack Jr., R.L. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.* **267:** 1268–1282.

Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein

sequences that fold into a known three-dimensional structure. *Science* **253:** 164–170.

Bruccoleri, R.E. and Karplus, M. 1990. Conformational sampling using high-temperature molecular dynamics. *Biopolymers* **29:** 1847–1862.

Carugo, O. and Pongor, S. 2001. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* **10:** 1470–1473.

Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., et al. 2002. Structural genomics: A pipeline for providing structures for the biologist. *Protein Sci.* **11:** 723–738.

Chothia, C. and Lesk, A.M. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196:** 901–917.

Chung, S.Y. and Subbiah, S. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure* **4:** 1123–1127.

Dahiyat, B.I. and Mayo, S.L. 1997. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci.* **94:** 10172–10177.

De Maeyer, M., Desmet, J., and Lasters, I. 1997. All in one: A highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.* **2:** 53–66.

Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins* **41:** 98–107.

Du, X., Choi, I.G., Kim, R., Wang, W., Jancarik, J., Yokota, H., and Kim, S.H. 2000. Crystal structure of an intracellular protease from *Pyrococcus horikoshii* at 2-Å resolution. *Proc. Natl. Acad. Sci.* **97:** 14079–14084.

Dunbrack, R.L. 2002. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12:** 431–440.

Eddy, S.R. 1995. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3:** 114–120.

———. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6:** 361–365.

Elofsson, A., Fischer, D., Rice, D.W., Le Grand, S.M., and Eisenberg, D. 1996. A study of combined structure/sequence profiles. *Fold. Des.* **1:** 451–461.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., and Bairoch, A. 2002. The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30:** 235–238.

Fetrow, J.S. and Skolnick, J. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281:** 949–968.

Fetrow, J.S., Siew, N., and Skolnick, J. 1999. Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* **13:** 1866–1874.

Fischer, D., Rice, D., Bowie, J.U., and Eisenberg, D. 1996. Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.* **10:** 126–136.

Fiser, A., Do, R.K., and Sali, A. 2000. Modeling of loops in protein structures. *Protein Sci.* **9:** 1753–1773.

Godzik, A. 1996. The structural alignment between two proteins: Is there a unique answer? *Protein Sci.* **5:** 1325–1338.

Gordon, D.B. and Mayo, S.L. 1999. Branch-and-terminate: A combinatorial optimization algorithm for protein design. *Struct. Fold. Des.* **7:** 1089–1098.

Gough, J. and Chothia, C. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30:** 268–272.

Griffiths-Jones, S. and Bateman, A. 2002. The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. *Bioinformatics* **18:** 1243–1249.

Haft, D.H., Loftus, B.J., Richardson, D.L., Yang, F., Eisen, J.A., Paulsen, I.T., and White, O. 2001. TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29:** 41–43.

Hargbo, J. and Elofsson, A. 1999. Hidden Markov models that use predicted secondary structures for fold recognition. *Proteins* **36:** 68–76.

Harrison, A., Pearl, F., Mott, R., Thornton, J., and Orengo, C. 2002. Quantifying the similarities within fold space. *J. Mol. Biol.* **323:** 909–926.

Hegyi, H. and Gerstein, M. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288:** 147–164.

Henikoff, S., Henikoff, J.G., and Pietrokovski, S. 1999. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15:** 471–479.

Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233:** 123–138.

Honig, B. and Nicholls, A. 1995. Classical electrostatics in biology and chemistry. *Science* **268:** 1144–1149.

Huang, J.Y. and Brutlag, D.L. 2001. The EMOTIF database. *Nucleic Acids Res.* **29:** 202–204.

Hubbard, T.J., Ailey, B., Brenner, S.E., Murzin, A.G., and Chothia, C. 1999.

SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **27:** 254–256.

Hwang, K.Y., Chung, J.H., Kim, S.H., Han, Y.S., and Cho, Y. 1999. Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nat. Struct. Biol.* **6:** 691–696.

Jacobson, M.P., Friesner, R.A., Xiang, Z., and Honig, B. 2002. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **320:** 597–608.

Jaroszewski, L., Rychlewski, L., Zhang, B., and Godzik, A. 1998. Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci.* **7:** 1431–1440.

Johnson, M.S., Overington, P.J., and Blundell, T.L. 1993. Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.* **231:** 735–752.

Jones, D.T. 1997. Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* **7:** 377–387.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. A new approach to protein fold recognition. *Nature* **358:** 86–89.

Jones, D.T., Tress, M., Bryson, K., and Hadley, C. 1999. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins* **Suppl 3:** 104–111.

Jones, S., and Thornton, J.M. 1997. Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272:** 133–143.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14:** 846–856.

Kasuya, A. and Thornton, J.M. 1999. Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.* **286:** 1673–1691.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299:** 499–520.

Kolinski, A., Betancourt, M.R., Kihara, D., Rotkiewicz, P., and Skolnick, J. 2001. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* **44:** 133–149.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235:** 1501–1531.

Landgraf, R., Xenarios, I., and Eisenberg, D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307:** 1487–1502.

Lazaridis, T. and Karplus, M. 2000. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10:** 139–145.

Lee, C. and Subbiah, S. 1991. Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* **217:** 373–388.

Leibowitz, N., Fligelman, Z.Y., Nussinov, R., and Wolfson, H.J. 2001. Automated multiple structure alignment and detection of a common substructural motif. *Proteins* **43:** 235–245.

Levitt, M. and Chothia, C. 1976. Structural patterns in globular proteins. *Nature* **261:** 552–558.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S., and Tsai, J. 1997. Protein folding: The endgame. *Annu. Rev. Biochem.* **66:** 549–579.

Lichtarge, O. and Sowa, M.E. 2002. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12:** 21–27.

Lichtarge, O., Bourne, J.R., and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257:** 342–358.

Looger, L.L. and Hellinga, H.W. 2001. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *J. Mol. Biol.* **307:** 429–445.

Luthy, R., Bowie, J.U., and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* **356:** 83–85.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316:** 139–154.

Madej, T., Gibrat, J.F., and Bryant, S.H. 1995. Threading a database of protein cores. *Proteins* **23:** 356–369.

Martin, A.C. and Thornton, J.M. 1996. Structural families in loops of homologous proteins: Automatic classification, modelling and application to antibodies. *J. Mol. Biol.* **263:** 800–815.

Martin, A.C., Orengo, C.A., Hutchinson, E.G., Jones, S., Karmirantzou, M., Laskowski, R.A., Mitchell, J.B., Taroni, C., and Thornton, J.M. 1998. Protein folds and functions. *Structure* **6:** 875–884.

Matsuo, Y. and Bryant, S.H. 1999. Identification of homologous core structures. *Proteins* **35:** 70–79.

Mirny, L.A. and Shakhnovich, E.I. 1999. Universally conserved positions in

protein folds: Reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291:** 177–196.

Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J.P. 1998. HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.* **7:** 2469–2471.

Murray, D. and Honig, B. 2002. Electrostatic control of the membrane targeting of C2 domains. *Mol. Cell* **9:**145–154.

Nayal, M., Hitz, B.C., and Honig, B. 1999. GRASS: A server for the graphical representation and analysis of structures. *Protein Sci.* **8:** 676–679.

Nicholls, A., Sharp, K.A., and Honig, B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11:** 281–296.

Nussinov, R. and Wolfson, H.J. 1991. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci.* **88:** 10495–10499.

Ogata, K., Ohya, M., and Umeyama, H. 1998. Amino acid similarity matrix for homology modeling derived from structural alignment and optimized by the Monte Carlo method. *J. Mol. Graph. Model* **16:** 178–189, 254.

Orengo, C.A. and Taylor, W.R. 1996. SSAP: Sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266:** 617–635.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* **5:** 1093–1108.

Ortiz, A.R., Strauss, C.E.M., and Olema, O. 2002. MAMMOTH (matching molecular molecular models obtained from theory): An automated method for model comparison. *Protein Sci.* **11:** 2606–2621.

Panchenko, A.R., Marchler-Bauer, A., and Bryant, S.H. 2000. Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* **296:** 1319–1331.

Pawlowski, K., Jaroszewski, L., Rychlewski, L., and Godzik, A. 2000. Sensitive sequence comparison as protein function predictor. *Pacific Symposium on Biocomputing* **5:** 42–53.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Pei, J. and Grishin, N.V. 2001. AL2CO: Calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17:** 700–712.

Petrey, D. and Honig, B. 2000. Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Sci.* **9:** 2181–2191.

Ponder, J.W., and Richards, F.M. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193:** 775–791.

Prlic, A., Domingues, F.S., and Sippl, M.J. 2000. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.* **13:** 545–550.

Rapp, C.S. and Friesner, R.A. 1999. Prediction of loop geometries using a generalized born model of solvation effects. *Proteins* **35:** 173–183.

Reddy, B.V., Li, W.W., Shindyalov, I.N., and Bourne, P.E. 2001. Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. *Proteins* **42:** 148–163.

Reva, B., Finkelstein, A., and Topiol, S. 2002. Threading with chemostructural restrictions method for predicting fold and functionally significant residues: Application to dipeptidylpeptidase IV (DPP-IV). *Proteins* **47:** 180–193.

Rice, D.W., Fischer, D., Weiss, R., and Eisenberg, D. 1997. Fold assignments for amino acid sequences of the CASP2 experiment. *Proteins* **Suppl 1:** 113–122.

Richardson, J.S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34:** 167–339.

Rost, B. 1997. Protein structures sustain evolutionary drift. *Fold. Des.* **2:** S19–S24.

———. 2002. Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318:** 595–608.

Rost, B., Schneider, R., and Sander, C. 1997. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270:** 471–480.

Russell, R.B., Sasieni, P.D., and Sternberg, M.J. 1998. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282:** 903–918.

Sali, A. and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234:** 779–815.

Samudrala, R. and Moult, J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275:** 895–916.

Sanchez, R. and Sali, A. 1997. Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.* **7:** 206–214.

Schmitt, S., Kuhn, D., and Klebe, G. 2002. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **323:** 387–406.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. 2000. SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28:** 231–234.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., and Kahn, D. 2002. ProDom: Automated clustering of homologous domains. *Brief Bioinform.* **3:** 246–251.

Sheinerman, F.B. and Honig, B. 2002. On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.* **318:** 161–177.

Shi, J., Blundell, T.L., and Mizuguchi, K. 2001. FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310:** 243–257.

Shindyalov, I.N. and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11:** 739–747.

———. 2000. An alternative view of protein fold space. *Proteins* **38:** 247–260.

Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34:** 82–95.

Sippl, M.J. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17:** 355–362.

———. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5:** 229–235.

Swindells, M.B., Orengo, C.A., Jones, D.T., Hutchinson, E.G., and Thornton, J.M. 1998. Contemporary approaches to protein structure classification. *Bioessays* **20:** 884–891.

Teichmann, S.A., Murzin, A.G., and Chothia, C. 2001. Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.* **11:** 354–363.

Todd, A.E., Orengo, C.A., and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307:** 1113–1143.

Tramontano, A. 1998. Homology modeling with low sequence identity. *Methods* **14:** 293–300.

Vasquez, M. 1996. Modeling side-chain conformation. *Curr. Opin. Struct. Biol.* **6:** 217–221.

Venclovas, C. 2001. Comparative modeling of CASP4 target proteins: Combining results of sequence search with three-dimensional structure assessment. *Proteins* **Suppl 5:** 47–54.

Vitkup, D., Melamud, E., Moult, J., and Sander, C. 2001. Completeness in structural genomics. *Nat. Struct. Biol.* **8:** 59–66.

Volz, K. 1999. A test case for structure-based functional assignment: The 1.2 Å crystal structure of the yjgF gene product from *Escherichia coli*. *Protein Sci.* **8:** 2428–2437.

Wallace, A.C., Laskowski, R.A., and Thornton, J.M. 1996. Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5:** 1001–1013.

Wallace, A.C., Borkakoti, N., and Thornton, J.M. 1997. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6:** 2308–2323.

Wilson, C.A., Kreychman, J., and Gerstein, M. 2000. Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297:** 233–249.

Xiang, Z. and Honig, B. 2001. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311:** 421–430.

Xiang, Z., Soto, C.S., and Honig, B. 2002. Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci.* **99:** 7432–7437.

Yang, A.S. and Honig, B. 2000a. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301:** 665–678.

———. 2000b. An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.* **301:** 679–689.

———. 2000c. An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.* **301:** 691–711.

Zarembinski, T.I., Hung, L.W., Mueller-Dieckmann, J.H., Kim, K.K., Yokota, H., Kim, R., and Kim, S.H. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Acad. Sci.* **95:** 15189–15193.